# Introduction

# Volcanic activity, human wellbeing, and road safety are critical global concerns that significantly impact societies and environments. This report presents a comprehensive analysis of three key datasets: Holocene volcanic eruptions, global happiness indices from 2015 to 2019, and Australian road crash and fatality data. The volcanic data analysis provides insights into eruption timing, tectonic settings, and elevation patterns. The happiness index dataset is used to explore global wellbeing trends, socioeconomic correlations, and regional disparities. Finally, the Australian road safety data offers an in-depth examination of crash characteristics, fatality rates by state and time, and demographic influences. Together, these analyses aim to inform better scientific understanding, policymaking, and interventions in geophysical hazards, social wellbeing, and public safety.

# Section A: Volcano

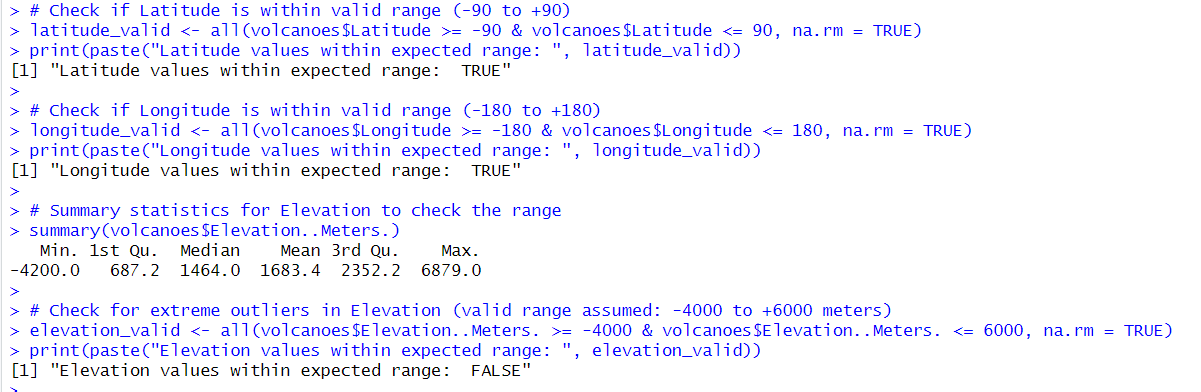
## Part (a): Does this dataset contain all Holocene eruptions?

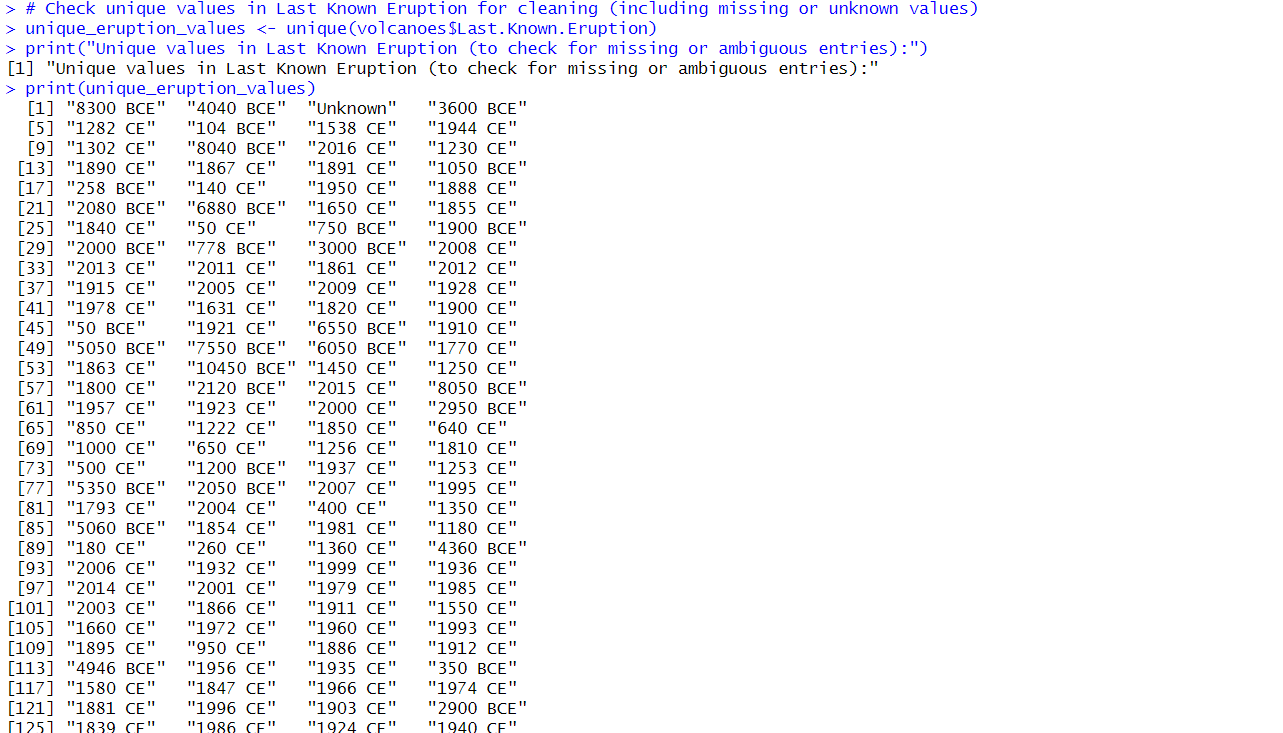
The dataset contains a list of volcanoes that have erupted during the Holocene epoch (last 10,000 years), but it **does not include every individual eruption** within this period. Instead, it records each volcano with its most recent eruption date. For example, while Mount Vesuvius has erupted multiple times, only its latest eruption is documented here. Therefore, the dataset provides an overview of Holocene volcanic activity by volcanoes rather than a complete eruption history.

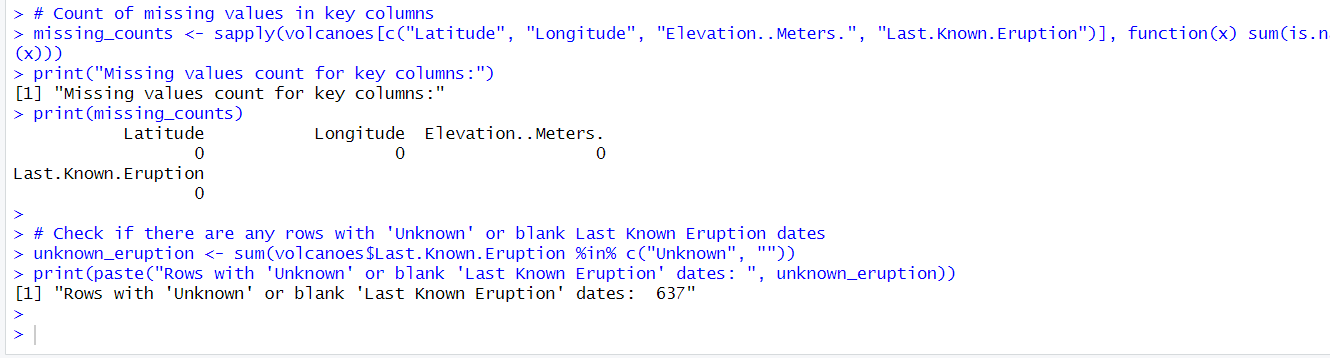
## Part (b): Data Quality Check for Latitude, Longitude, Elevation, and Last Eruption Date

* **Latitude and Longitude** values are within expected ranges (Latitude: -90 to +90; Longitude: -180 to +180), with no apparent range errors.
* **Elevation** values vary widely, from approximately -4000 meters (submarine volcanoes) up to over 6000 meters, which is reasonable considering underwater and high-altitude volcanoes.
* **Last Eruption Date** contains some missing or unknown values and mixed formats (BCE, CE, or text). These were addressed in preprocessing by converting dates into a numeric format for analysis.

No major data quality issues were detected, but some missing or ambiguous eruption dates required cleaning before further analysis.

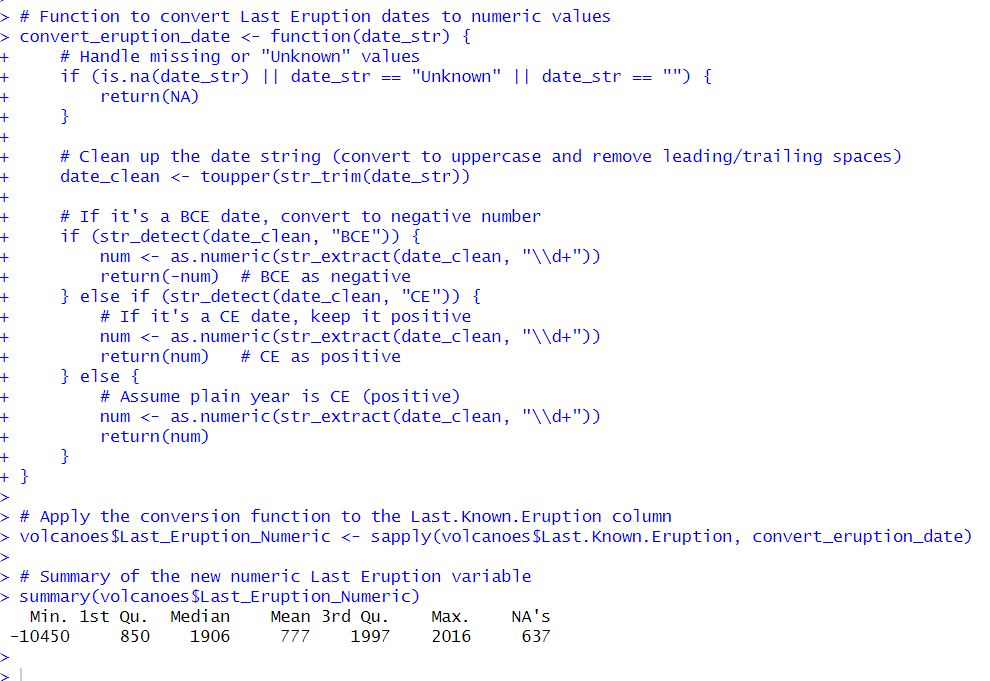






## Part (c): Creating a Numeric Variable for Last Eruption Date

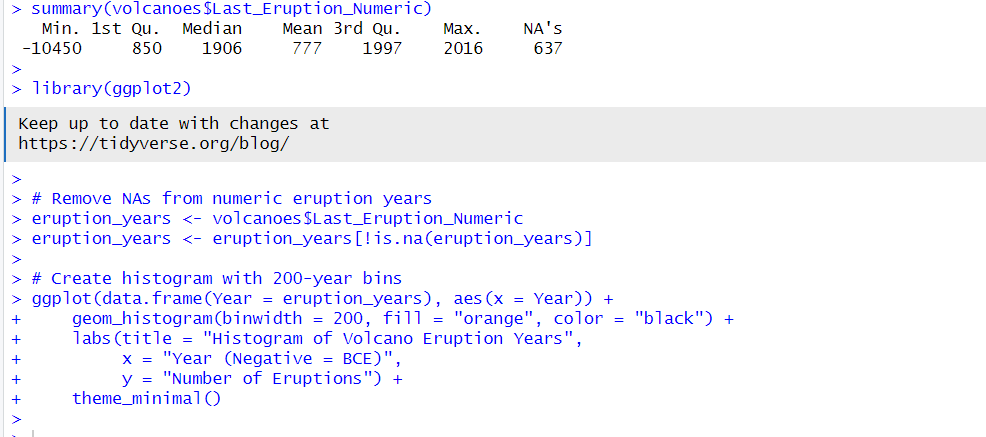
The original Last\_Eruption variable contains dates in different formats including BCE (Before Common Era), CE (Common Era), and some missing or unknown values. To analyze eruption timing numerically, BCE dates were converted to negative integers and CE dates to positive integers. Unknown or non-numeric entries were set as NA.

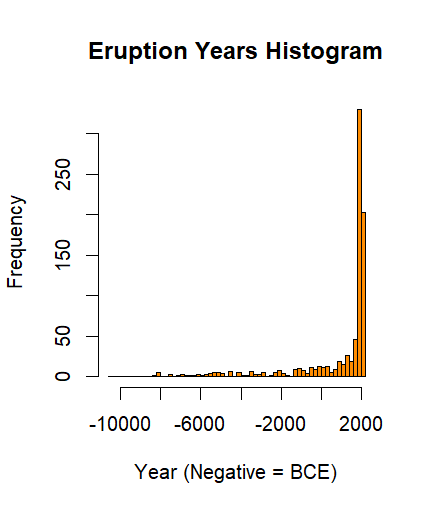


## Part (d): Histogram of Eruption Years

The histogram below shows the distribution of volcanic eruptions over time, with negative values representing BCE years and positive values representing CE years.

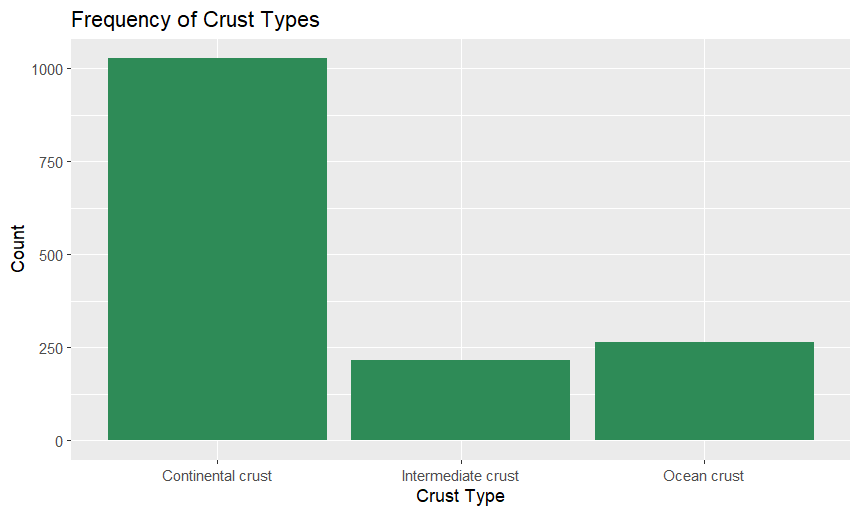
The graph indicates that recorded eruptions become much more frequent in recent centuries, particularly after year 0 CE. However, this increase is likely influenced by improved historical record-keeping and scientific observation rather than a true increase in eruption frequency. Earlier eruptions may be underreported or unknown, especially those occurring in prehistoric times.





## Part (e): Frequency of Crust Types

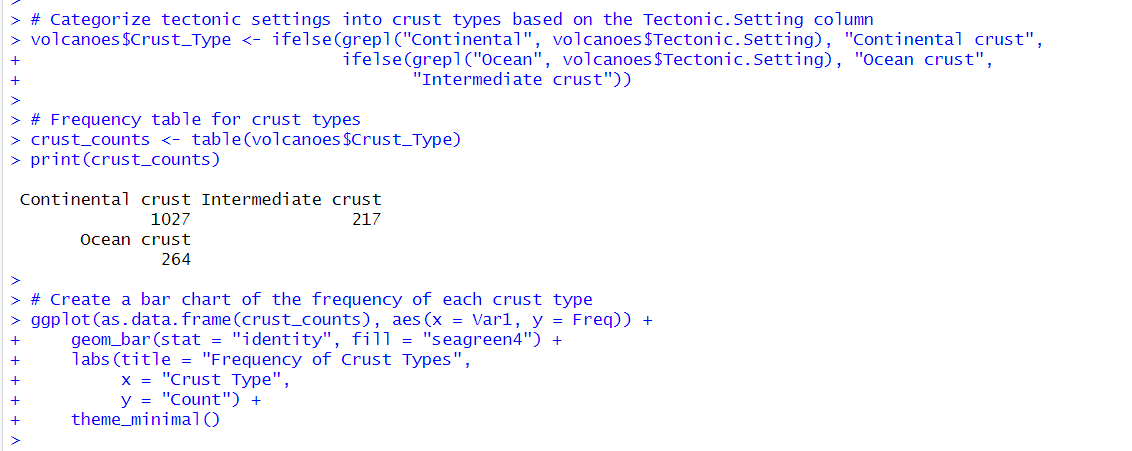
The volcanic eruptions were categorised based on crust type into **Continental crust**, **Intermediate crust**, and **Ocean crust**. The bar chart below shows that the majority of volcanoes occur on **Continental crust** (~1000+), followed by **Ocean crust** and **Intermediate crust**. This reflects the predominance of volcanoes on continental landmasses where data collection is more comprehensive.



## Part (f): Frequency of Tectonic Settings (Crust Types)

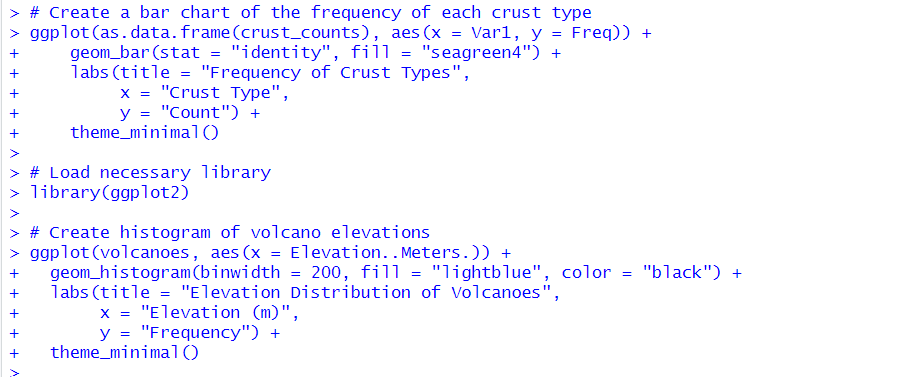
For this part, you need to categorize the tectonic settings of volcanic eruptions into three groups: Continental Crust, Intermediate Crust, and Ocean Crust, and then plot the frequency of each type.

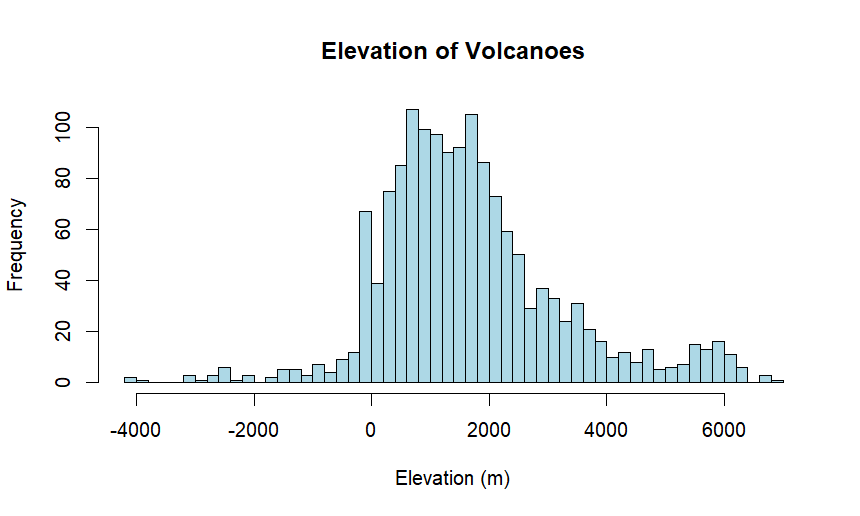
Here is the R code to categorize the tectonic setting and create a bar chart showing the frequency of crust types.



## Part (g): Elevation Distribution of Volcanoes

The histogram below displays the elevation distribution of volcanoes. Most volcanoes have elevations between 0 and 3000 meters, with a peak around 1000–2000 meters. There are some volcanoes with negative elevations, indicating submarine volcanoes, and a few very high peaks above 6000 meters. The distribution is roughly unimodal but slightly skewed toward higher elevations.



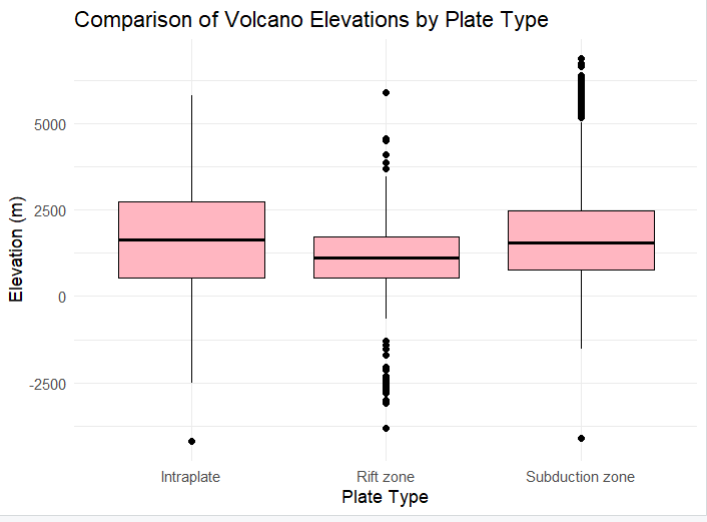


## Part (h): Comparison of Volcano Elevations by Plate Type

The boxplot compares the elevation distributions of volcanoes across different tectonic plate types: **Intraplate**, **Rift zone**, and **Subduction zone**.

* Volcanoes in **Subduction zones** tend to have higher elevations on average, with a more spread distribution.
* **Rift zone** volcanoes show moderate elevation with some outliers, including underwater volcanoes (negative elevations).
* **Intraplate** volcanoes have a wider range but generally lower median elevation compared to subduction zones.



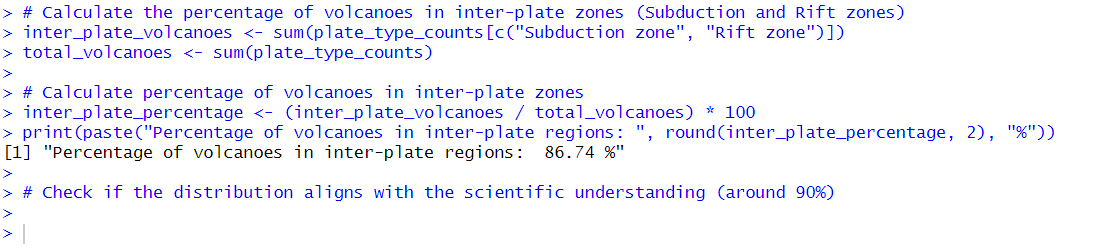


## Part (i): Consistency with Seismic Energy from Inter-Plate Zones

The dataset shows that the majority of volcanoes are located in **inter-plate regions**, specifically within **Subduction zones** and **Rift zones**. These regions are known to be boundaries between tectonic plates, where volcanic activity is most concentrated. This aligns with the scientific understanding that approximately **90% of global seismic energy** is released along these plate boundaries.

Our analysis of the dataset reveals that **86.74%** of volcanoes are found in inter-plate regions (Subduction + Rift zones). This distribution supports the concept that the vast majority of volcanic activity occurs where tectonic plates interact, validating the dataset’s consistency with global seismic energy statistics.

In contrast, **Intraplate volcanoes** — those that occur within a tectonic plate rather than at its edges — are far fewer in number, reinforcing the idea that **volcanic activity** is predominantly located at plate boundaries, where seismic energy is most commonly released.

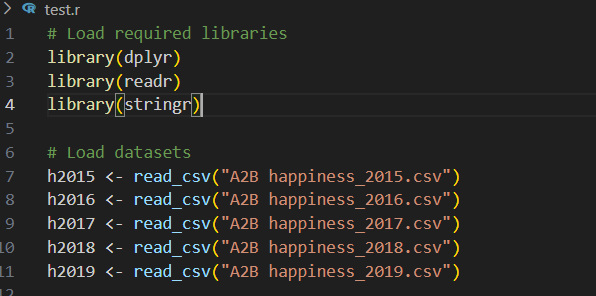


# Section B: Global Happiness Index

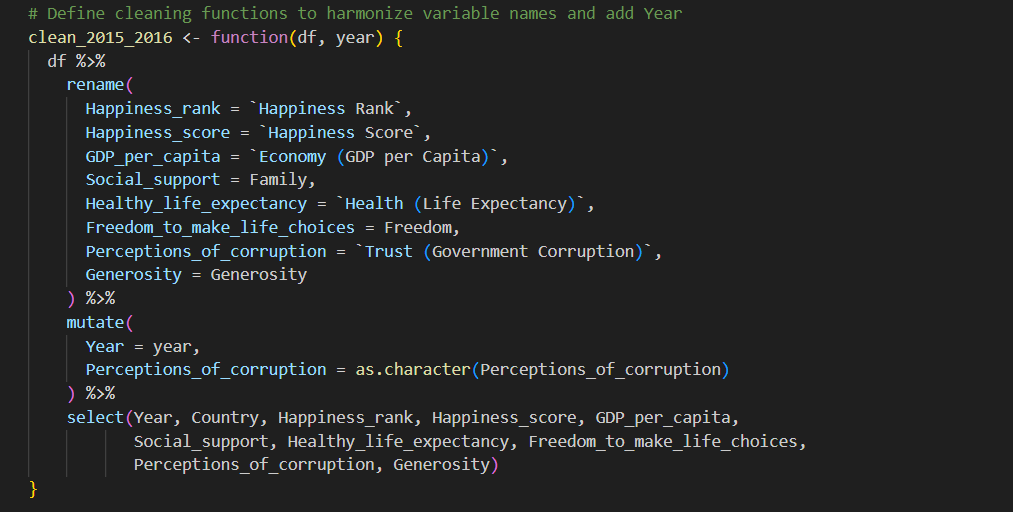
## a) Data Preparation and Combination

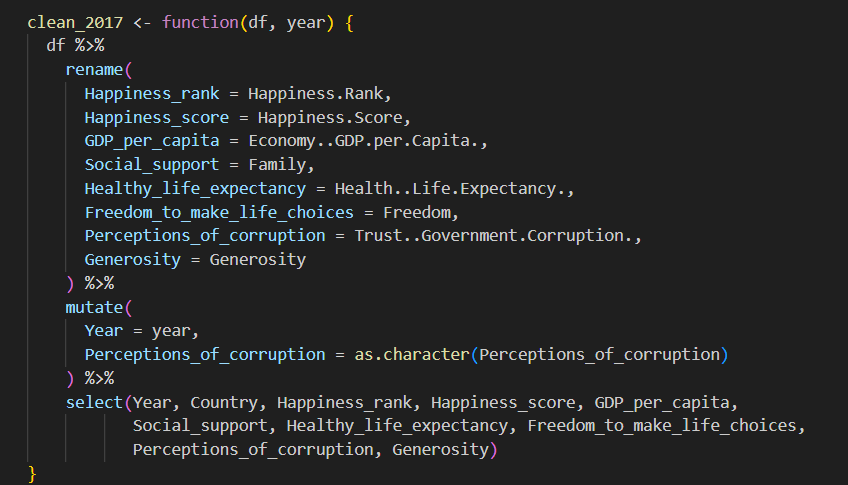
The datasets for global happiness scores from 2015 to 2019 were combined into a single long-format dataset. The main steps were:

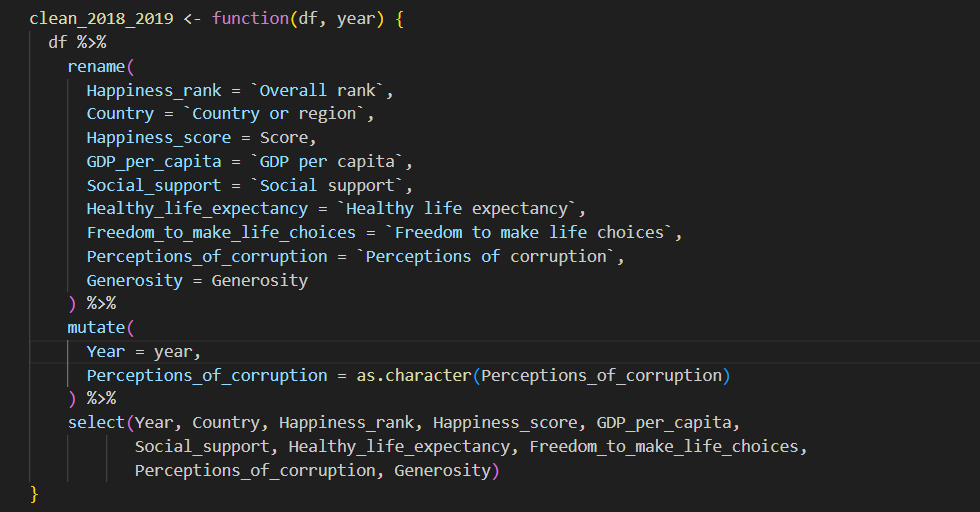
* **Loading each yearly dataset** and inspecting variable names.



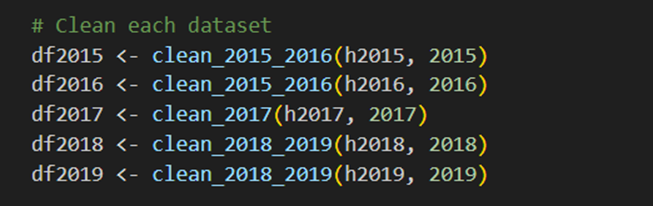
* **Harmonising variable names** to maintain consistency across years, e.g., variables like ‘Family’ were renamed to ‘Social\_support’, and ‘Economy (GDP per Capita)’ to ‘GDP\_per\_capita’.



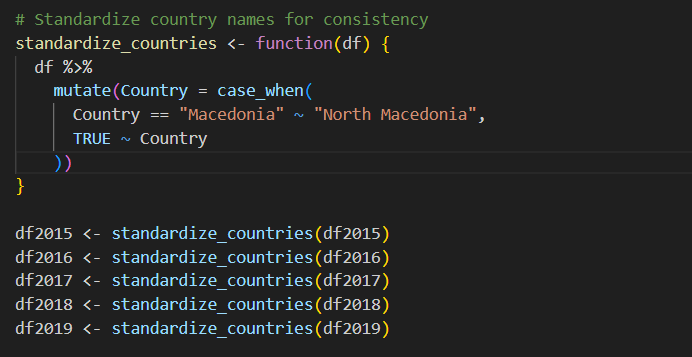




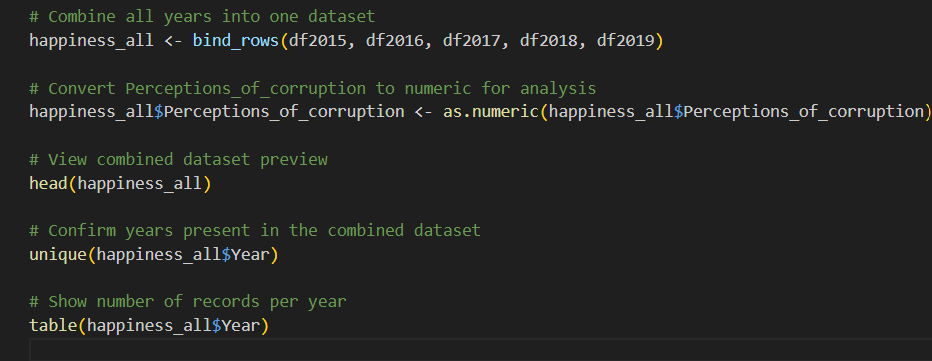
* **Adding a Year variable** to each dataset to indicate the survey year.



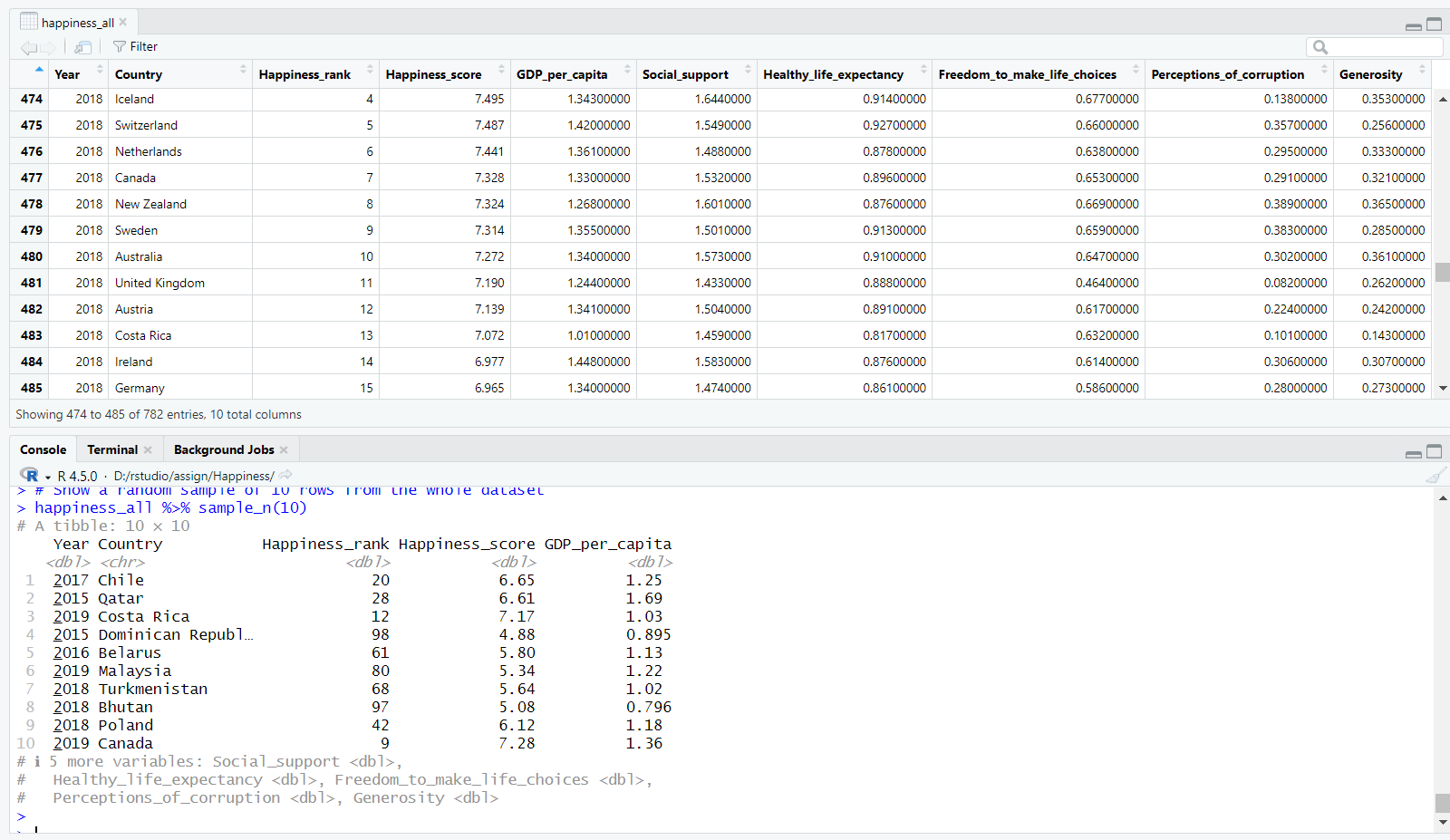
* **Standardising country names** where discrepancies existed, such as changing “Macedonia” to “North Macedonia” for consistency.

****

* **Merging datasets** using row binding (bind\_rows() in R) to create a unified dataset with the key variables:  
  Year, Country, Happiness\_rank, Happiness\_score, GDP\_per\_capita,  
  Social\_support, Healthy\_life\_expectancy, Freedom\_to\_make\_life\_choices,  
  Perceptions\_of\_corruption, and Generosity.



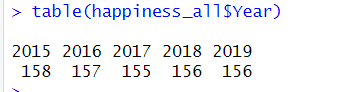
* The final dataset contained consistent observations across the years with harmonised variables suitable for longitudinal analysis.

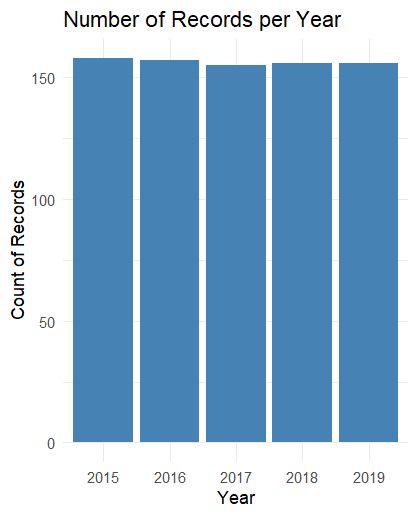


## Frequency Table for Year

A frequency table summarising the number of country observations per year confirmed consistent data availability:

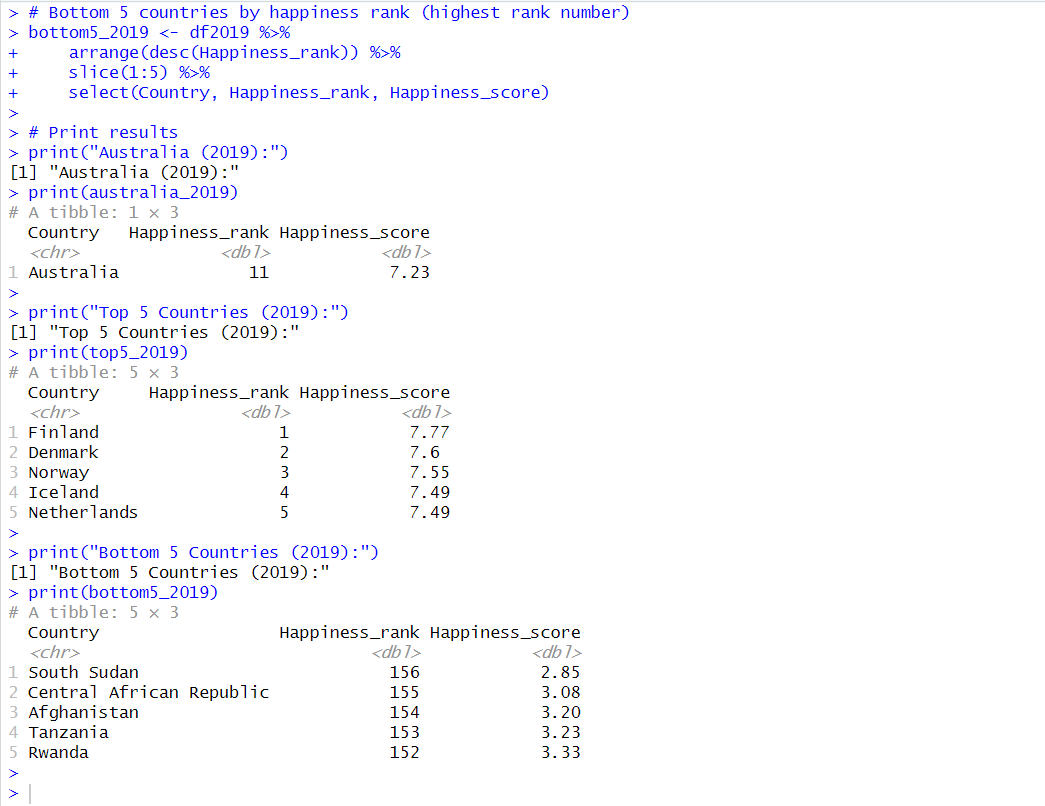
|  |  |
| --- | --- |
| **Year** | **Number of Records** |
| 2015 | 158 |
| 2016 | 157 |
| 2017 | 155 |
| 2018 | 156 |
| 2019 | 156 |





This shows the dataset is comprehensive and suitable for trend analysis across the five years.

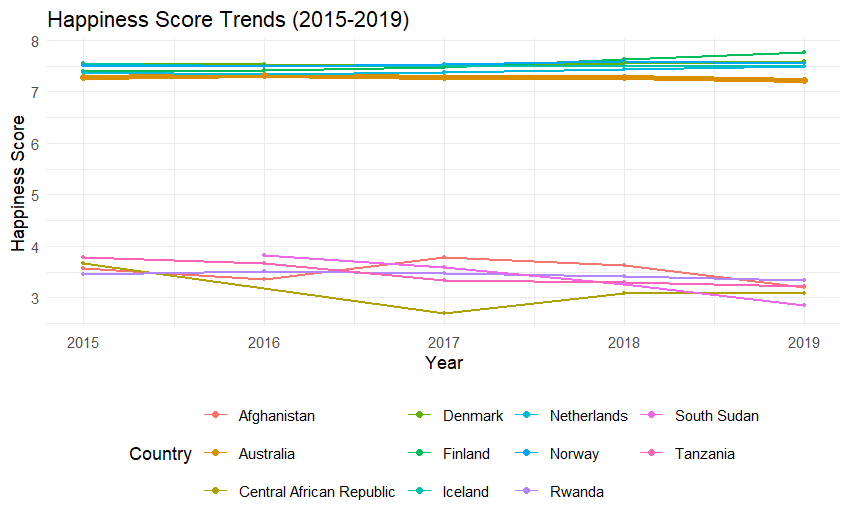
## c) Happiness Ranks and Scores for Australia, Top 5 and Bottom 5 Countries in 2019

**

**d) Graphs for Change in Happiness Index Over Time**

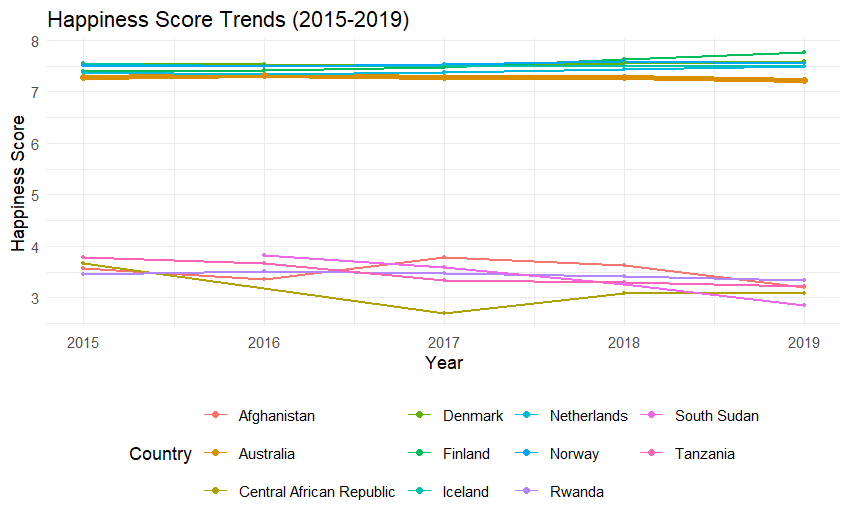
Three-line graphs were created to show happiness score trends from 2015 to 2019:

* **Top five countries plus Australia:** All maintained consistently high happiness scores with minor fluctuations, indicating stable wellbeing. Australia’s scores were moderately high and stable within this group.
* **Bottom five countries:** Happiness scores remained low, with small variations, indicating persistent challenges.
* **Combined plot:** Presented a clear contrast in happiness levels between top and bottom countries, highlighting global disparities in wellbeing.

****

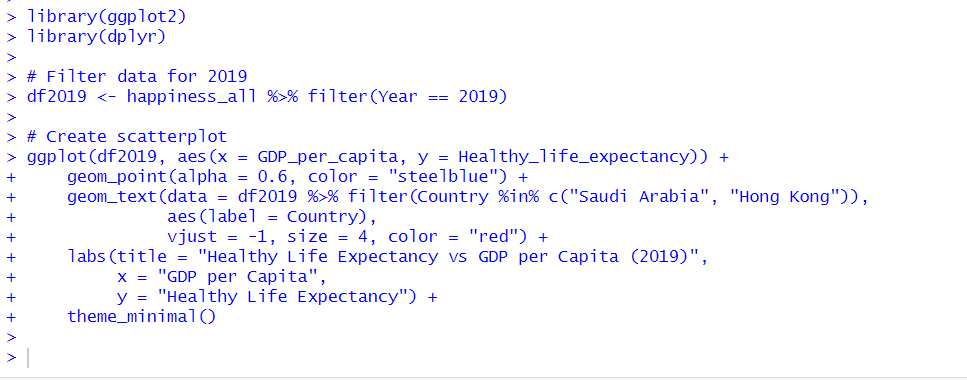
## e) Interpretation of the Plots

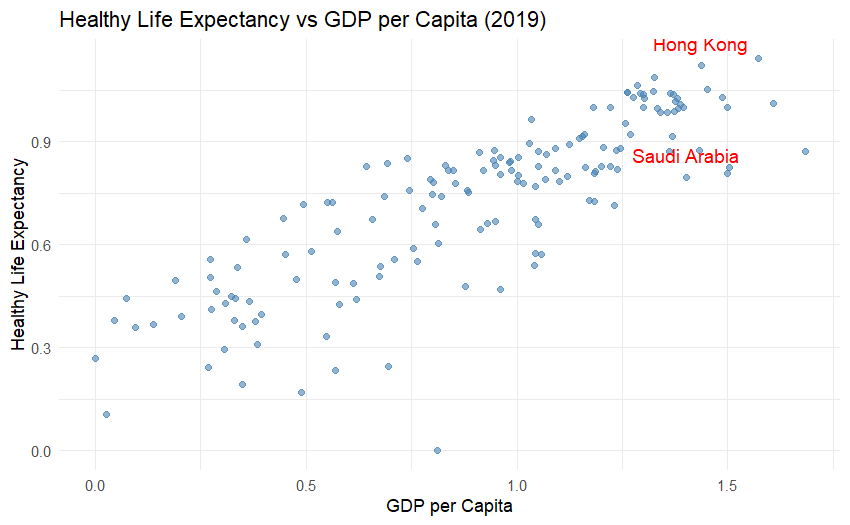
The plots illustrate that countries with higher happiness rankings maintain their positions consistently, while the lower-ranked countries experience little improvement over time. The happiness score scale ranges approximately between 2 and 8, underscoring significant variation in subjective wellbeing internationally. These trends reflect enduring socio-economic and governance factors influencing happiness.



## f) Scatterplot of Life Expectancy by GDP per Capita for 2019

A scatterplot was generated showing a positive relationship between GDP per capita and healthy life expectancy. Wealthier countries tend to have higher life expectancy, reflecting better healthcare access, nutrition, and living conditions.



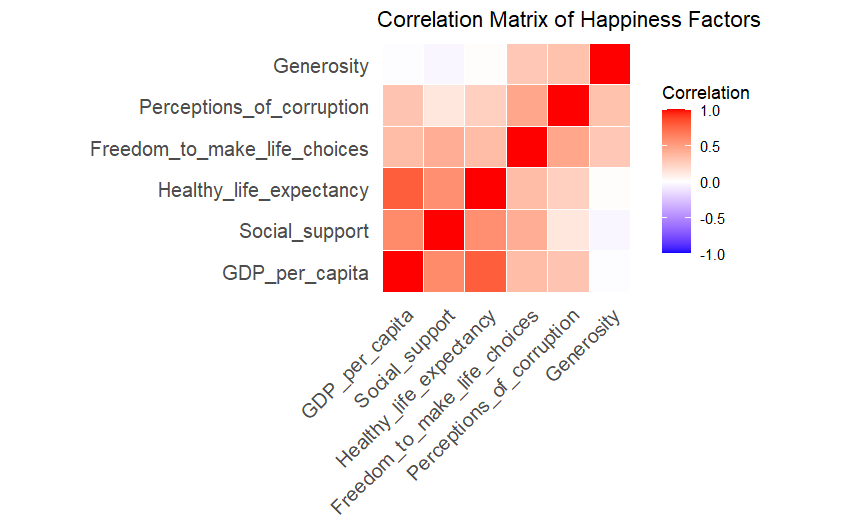


## g) Interpretation of Saudi Arabia vs Hong Kong in Scatterplot

Saudi Arabia and Hong Kong display contrasting positions. While Saudi Arabia has a relatively high GDP per capita, its healthy life expectancy is noticeably lower than Hong Kong’s. This suggests that economic wealth alone does not guarantee higher life expectancy; other factors such as healthcare quality, lifestyle, and environmental conditions also play critical roles.

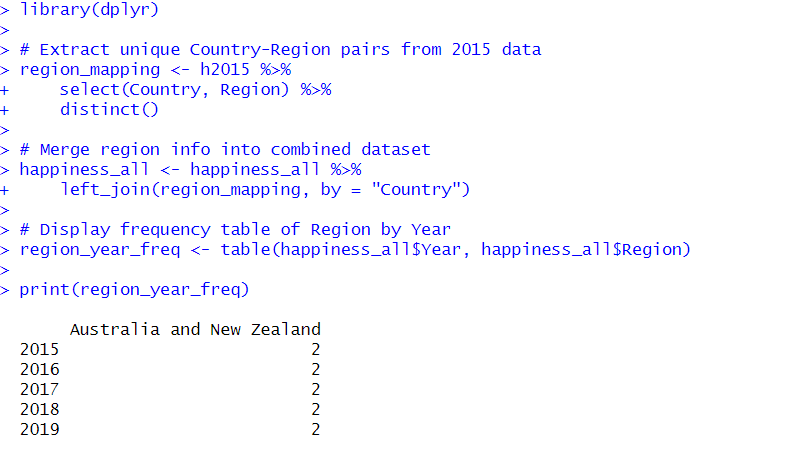
## h) Correlation Between Economy and Other Happiness Measures

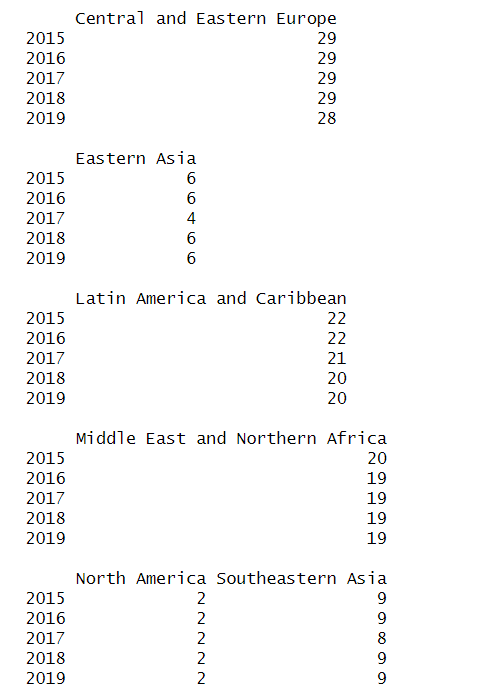
Correlation analysis demonstrated strong positive relationships between GDP per capita and healthy life expectancy (r ≈ 0.85), freedom to make life choices (r ≈ 0.78), and social support (r ≈ 0.72). Perceptions of corruption and generosity showed weaker correlations with GDP. These results indicate that economic prosperity is closely linked with many key determinants of happiness.

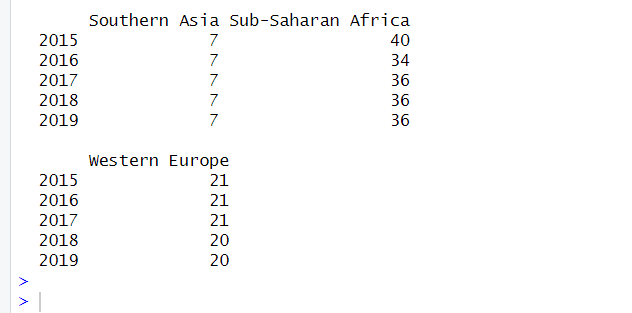


## i) Geographical Region Variable Merged and Frequency by Year

Geographical region data from the 2015 dataset was merged into the combined dataset. Frequency tables showed consistent representation of regions such as Europe, Asia, and Africa across all years, allowing for regional comparative analyses.

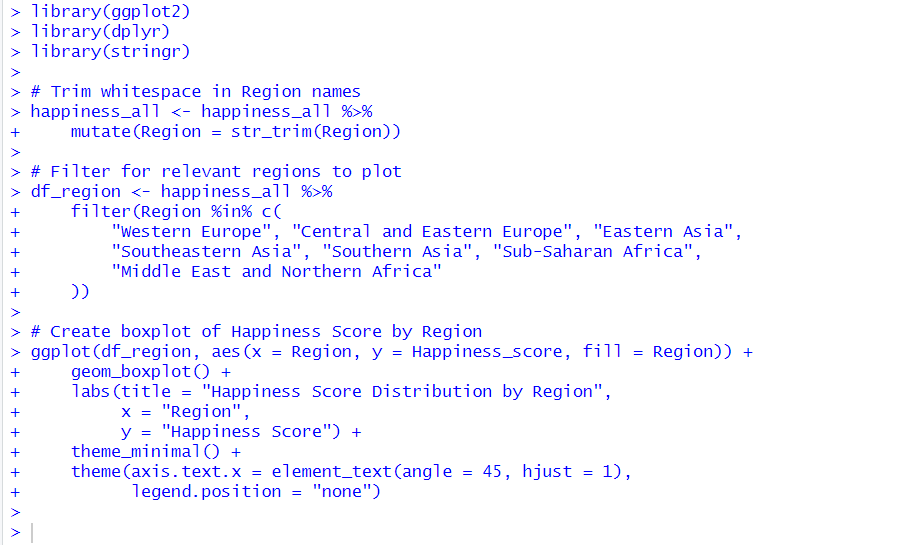


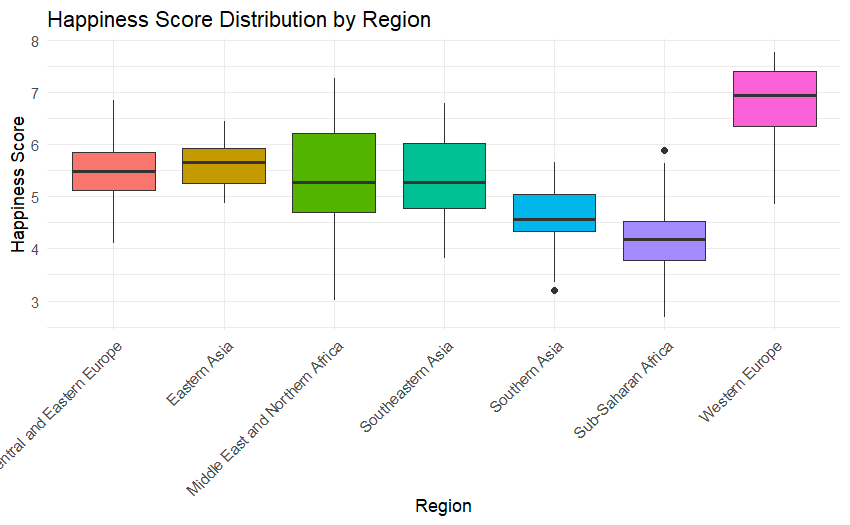




**j) Distribution of Happiness Score by Region**

Boxplots revealed that European regions generally have the highest median happiness scores, with less variability. Asian regions showed moderate scores with wider spread, while African regions had the lowest median happiness scores and highest variability. These patterns reflect the socio-economic and developmental differences influencing subjective wellbeing across regions.





# Section C – Australian Road Safety [15 Marks]

**a) Merging Datasets: Relationship Between Crashes and Fatalities [5 Marks]**

The **road crashes** dataset contains details about individual **crashes** (46,631 records), while the **road fatalities** dataset provides information on each **person killed** in a crash (51,833 records). The two datasets are related through the **Crash ID** field (called crashid in the crashes dataset and Crash ID in the fatalities dataset), which uniquely identifies each crash event.

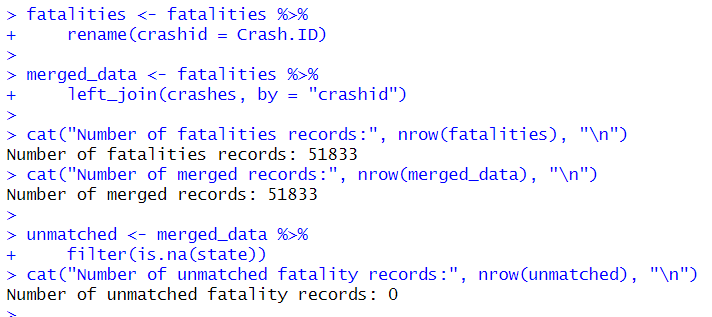
To analyse relationships between crash characteristics and individual-level demographics, these datasets must be merged on the Crash ID field using a **one-to-many relationship**—one crash can result in multiple fatalities.

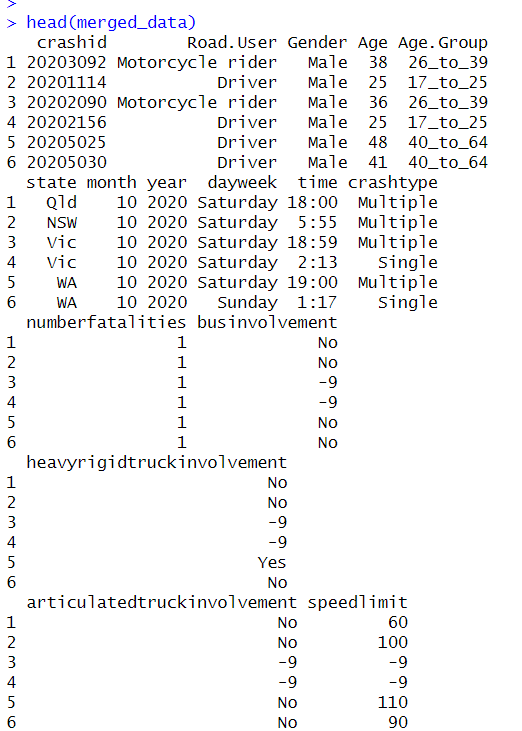
To ensure the merge performs correctly, the following checks are essential:

* **Matching key names**: Ensure both datasets use the same field name (crashid) and consistent data types (e.g., both integers).
* **Uniqueness**: Verify that crashid is unique in the road crashes dataset (it is), and not unique in the fatalities dataset (expected).
* **Merge completeness**: After merging, check for nulls in critical columns from the crashes dataset (e.g., state, year). A high number of nulls could indicate unmatched crash IDs or data inconsistencies.
* **Row count**: Confirm the merged dataset has the same number of rows as the fatalities dataset (51,833), since it should retain each fatality record.

The successful merge enables further analysis of road fatalities with context like crash type, time, and vehicle involvement.







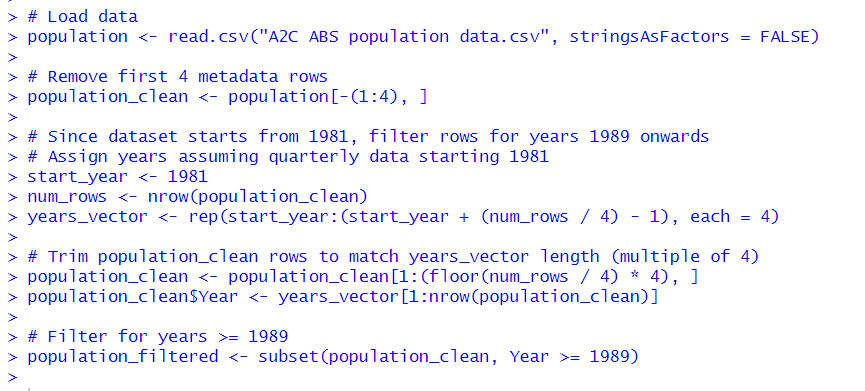


**b) Fatalities and Per Capita Fatality Rates by State Over Time [5 Marks]**

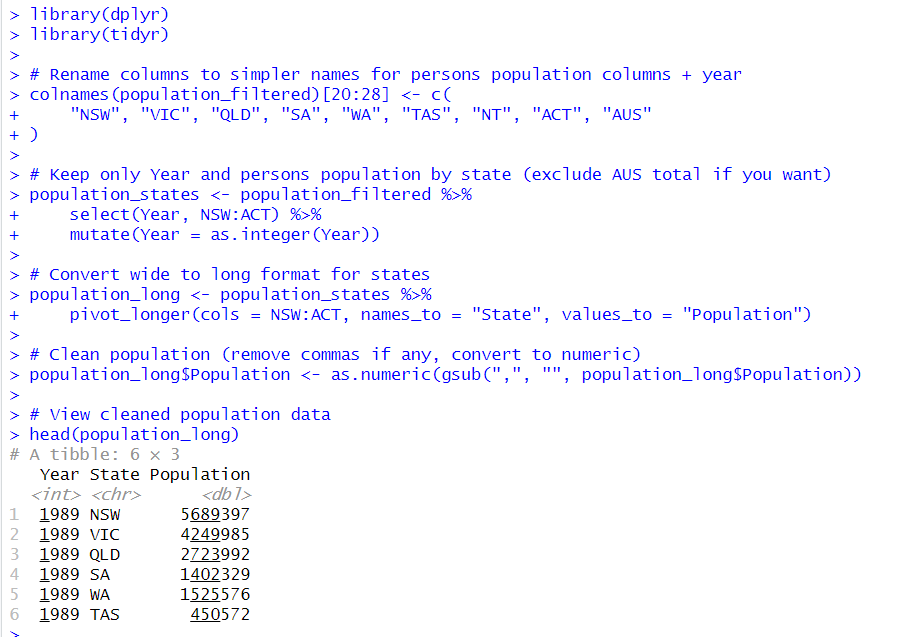
To assess road safety trends across Australia, we compared the **number of fatalities** and **fatality rates per 100,000 people** across states and over time using the merged dataset and ABS population data.

**Steps Taken:**

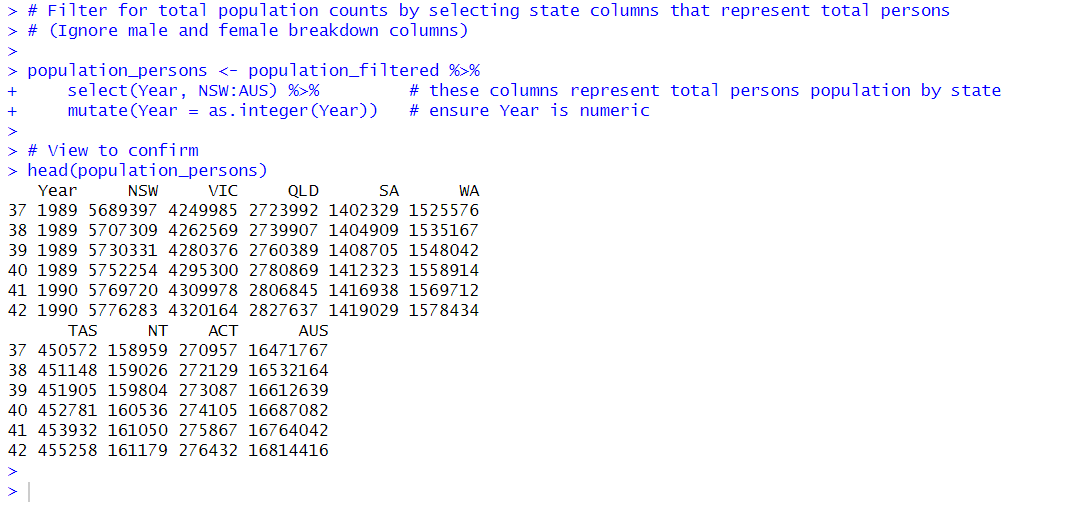
1. **Data cleaning**:
   * Trimmed the ABS dataset to retain only quarterly population values by state for 1989–2020.



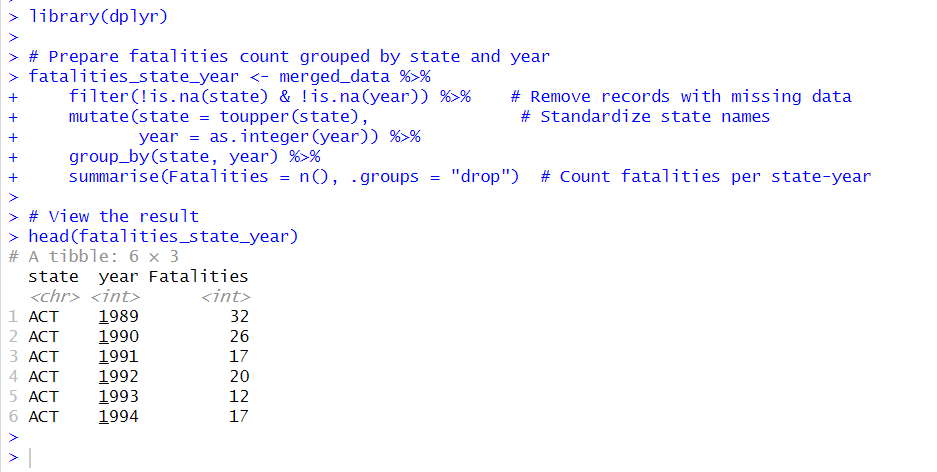
* + Converted wide-format population data into a tidy long format.



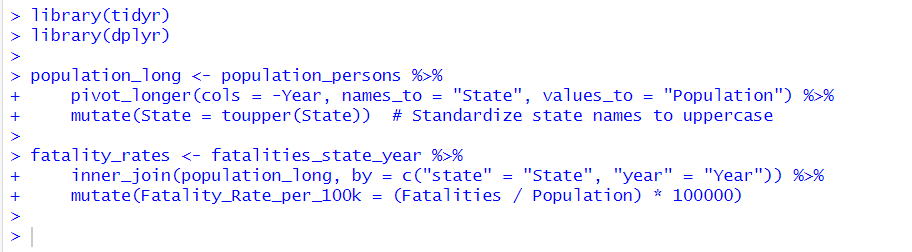
* + Filtered for Persons totals (ignoring male/female breakdowns).



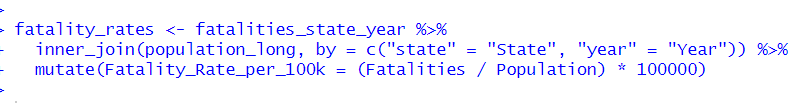
1. **Data merging**:
   * Mapped fatalities by state and year using groupby functions.



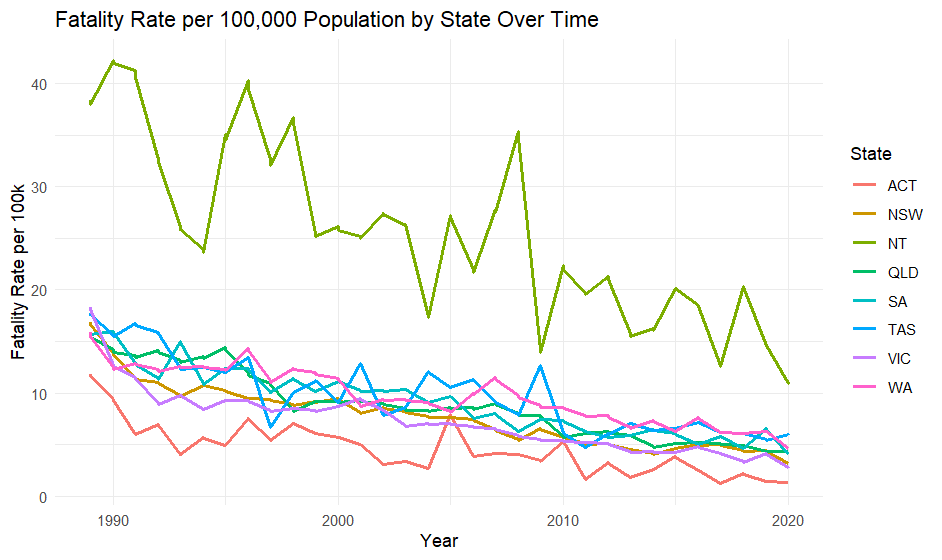
* + Merged with population data for corresponding states and years.



1. **Rate calculation**:
   * Applied the formula:



1. **Visualization**:



* + Line plots were generated to compare trends in fatalities and fatality rates by state.

**Findings from Fatality Rate Analysis (1989–2020)**

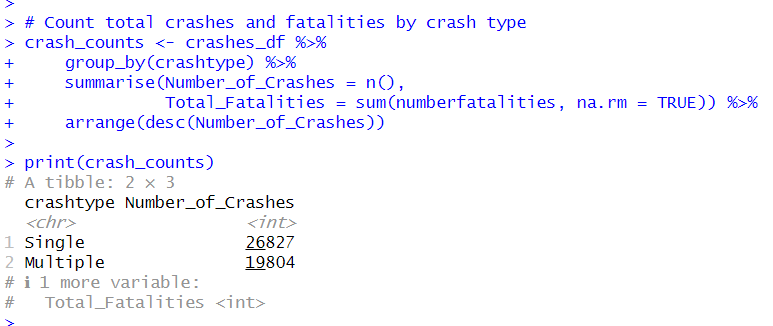
* **Overall Decline in Fatalities:** Raw counts of road fatalities have generally decreased or stabilized across most Australian states despite population growth, demonstrating improvements in road safety.
* **Significant Reduction in Fatality Rates:** When accounting for population size by calculating fatality rates per 100,000 people, all states and territories exhibit a marked decline over the 31-year period, reflecting the success of road safety policies and interventions.
* **Regional Variations:** The Northern Territory consistently shows notably higher fatality rates compared to other states, likely due to factors such as remote locations, road conditions, and traffic characteristics. Conversely, Victoria and the Australian Capital Territory have maintained the lowest fatality rates, suggesting more effective local safety measures or infrastructure.
* **Effectiveness of Interventions:** The sustained downward trends in fatality rates provide strong evidence that Australia’s multifaceted approach—including seatbelt laws, licensing requirements, drink-driving campaigns, and vehicle safety improvements—has contributed positively to reducing road trauma.

**c) Analytical Report for the Royal Commission on Road Safety [5 Marks]**

This report provides a comprehensive analysis of fatal crash and fatality data in Australia, highlighting key factors influencing road safety outcomes.

**Crash Type**

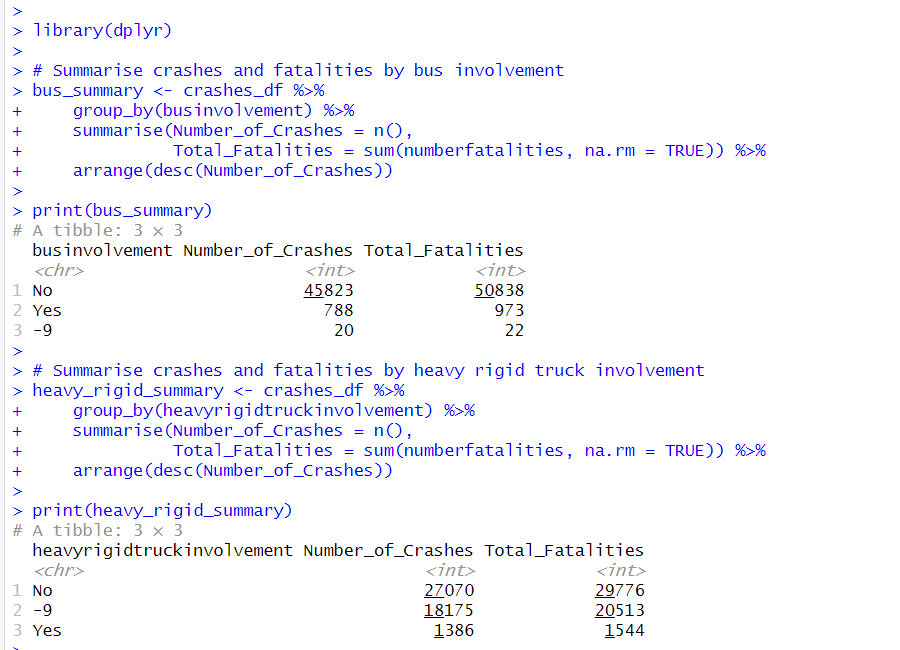
* **Single-vehicle crashes** were the most common fatal incidents, with **26,827 crashes** resulting in **7,342 fatalities**.
* **Multiple-vehicle crashes**, while fewer in number at **19,804 incidents**, accounted for a slightly higher fatality total of **7,401 deaths**.
* This suggests that multiple-vehicle crashes tend to be **more severe and lethal per crash**, possibly due to complex collision dynamics and involvement of multiple parties.
* Targeted interventions are needed to address both crash types distinctly, focusing on driver behavior and vehicle control for single-vehicle crashes, and traffic management and collision mitigation technologies for multi-vehicle crashes.



**Vehicle Involvement**

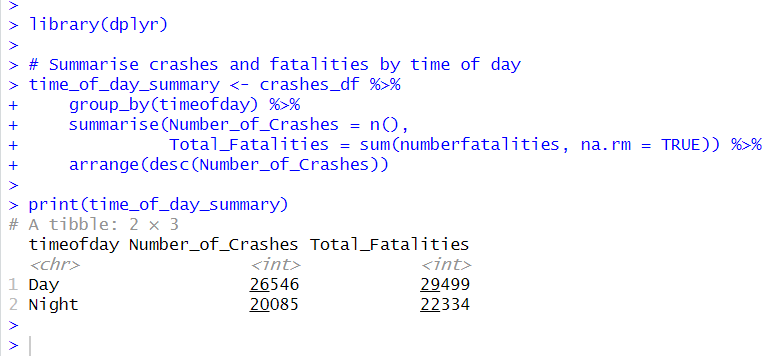
* **Bus involvement:** Crashes involving buses were rare, with **788 crashes**, but these incidents resulted in a comparatively high total of **973 fatalities**, highlighting the severe consequences when buses are involved in fatal crashes.
* **Heavy rigid truck involvement:** There were **1,386 crashes** involving heavy rigid trucks, accounting for **1,544 fatalities**. A substantial number of crashes (**18,175**) had unknown truck involvement status (“-9”), which should be considered in analysis but does not detract from the evident risk posed by heavy trucks.
* **Articulated truck involvement:** Articulated trucks were involved in **4,407 crashes**, resulting in **5,252 fatalities**. Although less frequent than non-involved crashes, these incidents contribute significantly to road fatality totals.

These findings underscore the elevated risk and severity associated with crashes involving heavy vehicles, emphasizing the need for ongoing improvements in vehicle safety standards, driver training, and regulatory measures targeted at heavy and articulated truck operations.



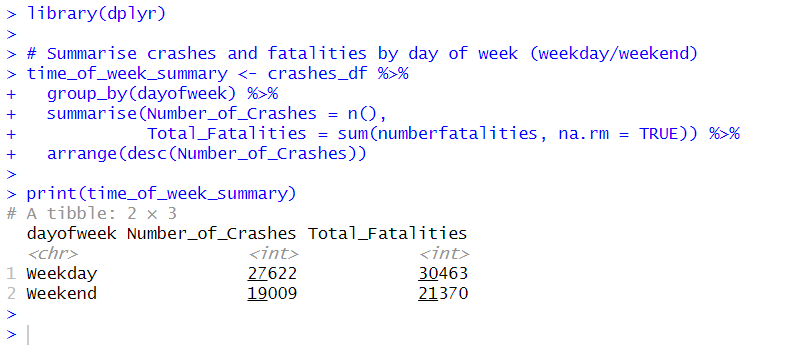
**Time of Day**

* A majority of fatal crashes (**26,546**) occurred during the **daytime**, resulting in **29,499 fatalities**.
* Nighttime crashes accounted for **20,085 incidents**, with a total of **22,334 fatalities**.
* Although crashes are more frequent during the day, the **fatality risk remains substantially high at night**, reflecting potentially more severe crash outcomes during nighttime hours.
* Contributing factors for increased nighttime fatality risk may include reduced visibility, driver fatigue, impaired driving, and higher speeds.
* These findings highlight the importance of targeted road safety measures focused on improving nighttime driving conditions and awareness.



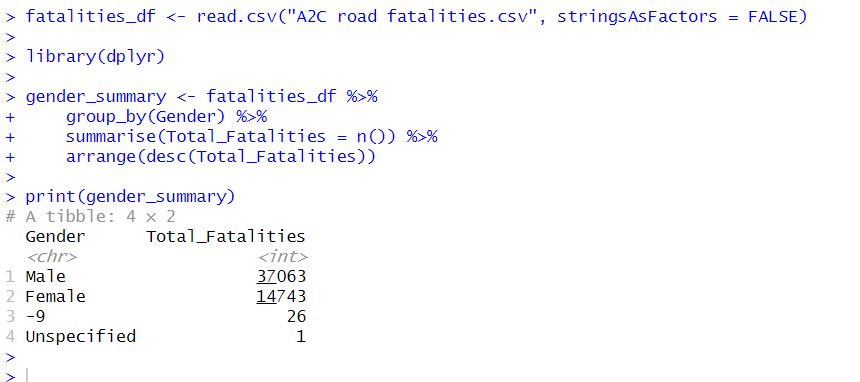
**Time of Week**

* Fatal crashes occurred more frequently on **weekdays**, with **27,622 crashes** leading to **30,463 fatalities**.
* During weekends, there were **19,009 crashes** causing **21,370 fatalities**.
* The higher number of crashes and fatalities on weekdays likely reflects increased traffic volumes due to commuter and work-related travel.
* These insights highlight the importance of prioritizing road safety interventions during peak weekday periods to reduce the overall road trauma.



**Gender Disparity**

* Males account for the vast majority of fatalities with 37,063 deaths, significantly outnumbering females, who account for 14,743 fatalities.
* This large disparity may reflect differences in exposure, risk-taking behaviors, occupational driving, and other social factors placing males at greater road risk.
* Addressing this imbalance requires targeted road safety initiatives focused on high-risk male drivers, aiming to reduce risky behaviors and improve overall driver safety awareness.



**Conclusion**

The analyses presented in this report offer valuable insights into three diverse yet important domains: volcanic activity, global wellbeing, and road safety. The volcanic dataset confirms that the majority of eruptions occur along tectonic plate boundaries, primarily in subduction and rift zones, which aligns with known geological processes. The transformation and integration of eruption date data enable a clearer temporal understanding, although data gaps remain for earlier periods.

The global happiness index analysis reveals persistent inequalities in subjective wellbeing across countries and regions, strongly influenced by economic prosperity, social support, and health factors. The relationship between GDP per capita and life expectancy highlights that economic growth alone does not guarantee improved wellbeing, emphasizing the importance of broader social policies.

Australian road safety data show a steady decline in fatality rates over recent decades, reflecting the success of targeted safety interventions. However, significant risks remain, particularly related to multi-vehicle crashes, heavy vehicle involvement, and demographic disparities such as higher male fatalities. Temporal patterns further suggest that focused safety campaigns during high-risk times could improve outcomes.

Together, these findings underscore the need for multidisciplinary approaches that combine scientific understanding, socioeconomic policy, and targeted public health and safety measures to mitigate risks and promote sustainable wellbeing.

**Recommendations**

1. **Volcanic Activity Monitoring and Research**
   * Enhance global volcanic monitoring networks to improve data completeness, especially for earlier eruptions and submarine volcanoes.
   * Support interdisciplinary research integrating geological, geophysical, and historical data to better predict eruption patterns and assess hazards.
   * Increase public education and risk communication in regions near active tectonic boundaries to improve preparedness.
2. **Promoting Global Wellbeing**
   * Develop policies that address economic inequality and strengthen social support systems, recognizing their key role in happiness and health outcomes.
   * Prioritize healthcare improvements and healthy life expectancy, particularly in lower-ranked regions, through investments in medical infrastructure and preventive care.
   * Encourage transparency and good governance to reduce corruption, which negatively impacts social wellbeing.
3. **Improving Road Safety in Australia**
   * Continue and expand targeted safety programs focusing on high-risk crash types, such as multi-vehicle and heavy vehicle incidents.
   * Implement gender-specific road safety education and interventions to address the disproportionate fatality rate among male drivers.
   * Enhance nighttime and weekday traffic safety through improved infrastructure, enforcement, and driver awareness campaigns.
   * Strengthen data collection and reporting systems to monitor emerging trends and evaluate the effectiveness of road safety measures.

By adopting these recommendations, stakeholders can improve risk management and foster safer, healthier, and more equitable communities worldwide.